B. TECH. PROJECT REPORT

On

Attacking Semantic Segmentation Models

BY

Mohit Nathrani
Electrical Engineering

Mahesh Kumar
Computer Science and Engineering

DISCIPLINE OF COMPUTER SCIENCE AND ENGINEERING
And ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY INDORE

December 2019

Attacking Semantic Segmentation Models

A PROJECT REPORT

Submitted in partial fulfillment of the requirements for the award of the degrees

OF **BACHELOR OF TECHNOLOGY in**
COMPUTER SCIENCE AND ENGINEERING
And
ELECTRICAL ENGINEERING

Submitted by: **Mohit Nathrani and**
Mahesh Kumar

Guided by: **Dr. Puneet Gupta, Computer Science, and Engineering**
Assistant Professor, IIT Indore

INDIAN INSTITUTE OF TECHNOLOGY INDORE
December 2019

CANDIDATE'S DECLARATION:

We hereby declare that the project entitled "Attacking Semantic Segmentation Models" submitted in partial fulfillment for the award of the degree of Bachelor of Technology completed under the supervision of Dr. Puneet Gupta, Computer Science, and Engineering Assistant Professor, IIT Indore is an authentic work.

Further, I/we declare that I/we have not submitted this work for the award of any other degree elsewhere.

Signed:

Mohit Nathrani (160002030)

Mahesh Kumar (160001033)

Certificate

It is certified that the above statements made by the student are correct to the best of my knowledge.

Supervisor

Dr. Puneet Gupta,
Assistant Professor,
Computer Science and Engineering
Indian Institute of Technology Indore

Date:

Preface

This report on "Attacking Semantic Segmentation Models" is prepared under the guidance of Dr. Puneet Gupta.

I have tried to present the detailed concept of different methods to attack Neural Networks. This report is the explanation of all the Attack techniques and is shown diagrammatically. As well as the algorithm for generating adversarial attack is present on our GitHub profile. For better understanding and visualization of our concept, we have added pictures. We have gone through step by step procedure for generating the final output. To conclude the report, a comparison of our work with the existing method is included.

Mohit Nathrani
4th Year B.Tech
Discipline of Electrical Engineering
IIT Indore

Mahesh Kumar
4th Year B.Tech
Discipline of Computer Science and Engineering
IIT Indore

Acknowledgments

I wish to thank Dr. Puneet Gupta for his kind support and valuable guidance.

It is his help and support, due to which I became able to complete the design and technical report.

Without his support, this report would not have been possible. It was only possible because of his enthusiasm, beforehand schedule, knowledge, and sincerity towards me and my work to produce a better result. I wouldn't have achieved my goals without his encouragement at each step.

This study has indeed helped us to explore more knowledgeable avenues related to this topic and we are sure it will help us in the future.

Mohit Nathrani
4th B.Tech
Discipline of Electrical Engineering
IIT Indore

Mahesh Kumar
4th B.Tech
Discipline of Computer Science and Engineering
IIT Indore

Abstract

Attacking Semantic Segmentation Models

Deep Learning has won almost all the recent competitions on image classifications and by now performs better than humans at recognizing objects in an image. In fact, self-driving car research is no longer available to only large companies. Startups are entering the market thanks to how cheap and powerful deep learning is for such systems. Even after the huge success of deep neural networks, their applications are limited because these networks can be fooled by adversarial examples.

Semantic segmentation is a fundamental building block of machine learning consisting of three basic steps of object detection, shape recognition, and classification. It is the most crucial part of autonomous driving, web security, image detection, and other computer vision task, but in contrast to image classification tasks, only very limited studies are available for attacking semantic segmentation networks. The existing semantic segmentation attacks use cross-entropy as the loss function. However, the success of the attack is generally measured using Intersection-Over- Union (IoU), which is non-differentiable. This gap between performance measure and loss function gave us the motivation of using a neural network as an approximation of IoU function. Then instead of using cross-entropy, this surrogate loss function can be used in any attack. Experiments and some results on VOC2012 (publicly available dataset) using state-of-the-art semantic segmentation network architectures are mentioned in this report.

Contents

List of Figures

Chapter 1

Introduction

1.1 Semantic Segmentation

Image segmentation is a computer vision process in which we label specific pixels of an image based on content in it. The goal of semantic image segmentation is to give a class to each pixel of an image with a respective class of which they belong. As shown in Fig 1 each pixel is labelled as a bicycle, person or background. Semantic segmentation is a fundamental building block of machine learning consisting of three basic stages of object detection, shape recognition, and classification. It is the most crucial part of autonomous driving, web security, image detection, and other computer vision tasks.

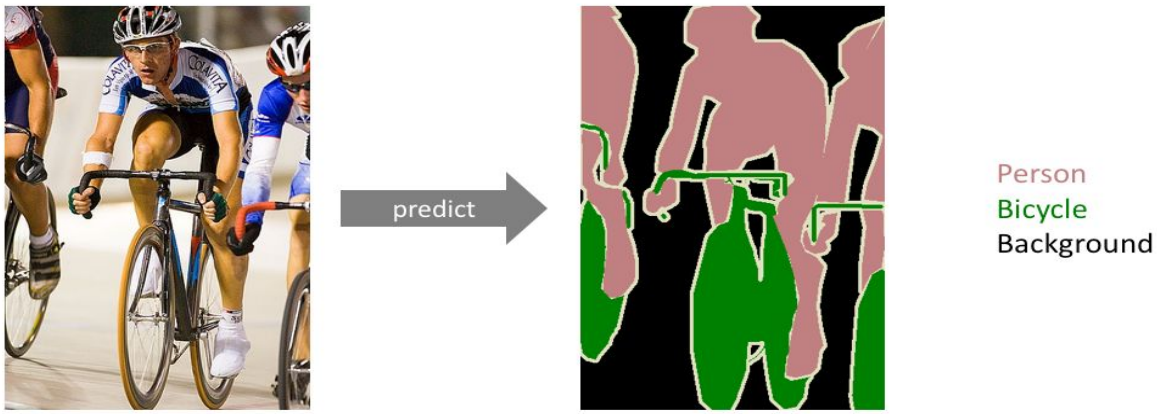


Fig 1. Semantic segmentation example [1]

1.2 Adversarial Attack

An adversarial attack consists of subtle modifications in a given image in such a way that the changes are almost unrecognizable from a human eye. This new modified image is called an adversarial image, and when fed into a classifier is misclassified, whereas the original one correctly classified. A hypothetical example is shown in Fig 2 where a is fed into an image classifier network. For original image is predicting that image contains a fish with very high confidence of 90%, but a generated adversarial example incorrectly predicts it to be a cat.

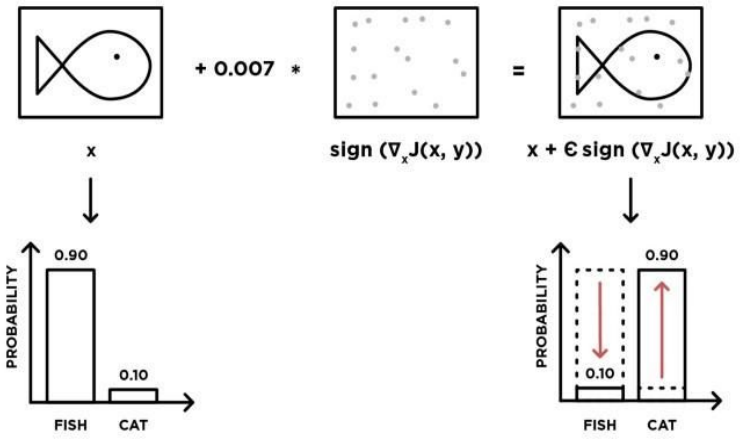


Fig 2. Adversarial attack example [2]

Mathematically, the generation of an adversarial example *x'* for an input image *x* can be modelled as an optimization problem.

$$
\begin{aligned}
min \quad & \|x' - x\| \\
s.t. \quad & f(x') = l', \\
& f(x) = l, \\
& l \neq l', \\
& x' \in [0, 1]
\end{aligned}
$$

where *l* and *l'* denote the label of *x* and *x'*. [3]

1.2.1 Real-World Adversarial Example

- We can now generate an adversarial eyeglass that when printed and used with a real eyeglass frame can fool face recognition models. As shown in Fig 3 a researcher designed an adversarial frame and now using a frame of this kind any person can classify himself as Milla Jovovich, an American actress [4]. This powerful technique can be used to generate a frame for any targeted person to fool any face recognition model. Criminal identification, advertising, facial biometrics and systems of this kind are vulnerable to this attack.

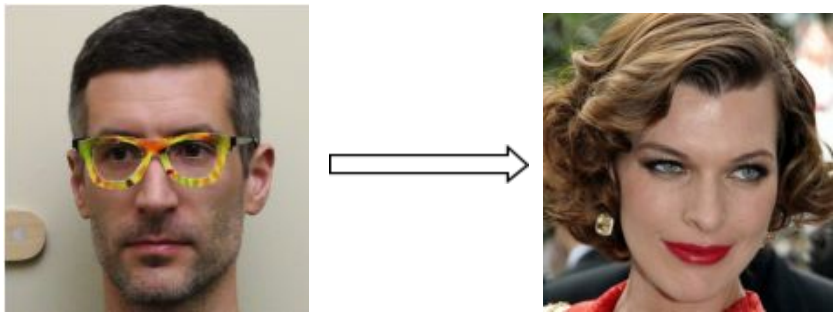


Fig 3. Example of a random person pretending to be Milla Jovovich using the eyeglass frame [4]

- A custom specially designed printed license plate as shown in Fig 4 can be made so that it can be incorrectly classified by any existing traffic speed camera even when it looks perfectly normal. If we put perturbation on a certain position in the number plate then any automatic toll camera system or number license-plate recognition system which detects the vehicle can be fooled. Since the technique used for generation is a black-box attack, it will fool almost all LPR systems.

Fig 4. Example of custom number plate with perturbation to confuse any LPR traffic camera system.

- An adversarial stop sign carefully generated can always be misclassified as a speed limit sign or any targeted sign, even when viewed from different angles while travelling [5]. A successful adversarial example of this where stop sign gets classified as speed limit 45 is shown in Fig 5 The exploitation of this type of attack can cause severe accidents and a lot of security issues. For example, any criminal person can stop any person's car by using a special kind of sticker on a traffic sign and can execute dangerous criminal activities.

Fig 5. Example of a stop sign with physical perturbation applied to fool classifier of self-driving cars which classifies it to speed limit 45 sign [5].

1.3 IoU

Intersection over Union (IOU) is a parameter for evaluation or measurement of the accuracy of object detection on a particular dataset. Intuitively IoU is easy to understand, a score of zero means no overlap of predicted result and ground truth. A score of one means the exact match of prediction and ground truth We frequently see this evaluation metric used in semantic segmentation, object detection, and other computer vision tasks. Mathematical and pictorial definition of IoU as shown in Fig 6.

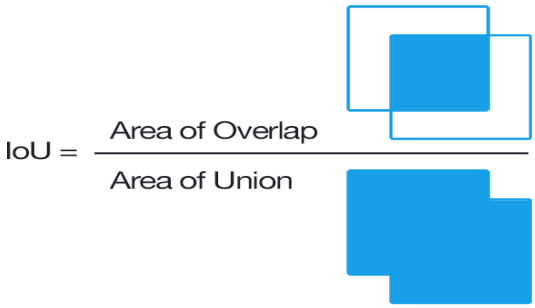


Fig 6. Mathematical and pictorial definition of IoU [6].

1.4. Cause of Adversarial Examples

The reason for the existence of adversarial examples is still not completely clear. Initially, researchers believed the reason is the over-fitting or under-regularization of the model which leads to the insufficient generalization ability of any neural network models shown in [7]. Then, others considered that adversarial examples exist because of the extreme nonlinearity of the neural networks as shown in [8]. However, a researcher added slight perturbations to the input to a linear model that had enough dimensions and demonstrated that model effectiveness of

defending from adversarial attacks was not necessarily improved. He believes that the reason for existence for adversarial examples is the linear nature in high dimensional space. So, perturbation for one-dimension of each input will not affect the overall prediction of the classifier, while small perturbations to all-dimensions of the inputs will lead to very significant change as shown in [9].

1.5 Need of Defense

Since adversarial examples exist for artificial intelligence systems, these vulnerabilities limit the applications and expansion of neural networks in sensitive security fields. The result of an attack on real-world applications can be very severe. For example, if someone modifies the traffic sign so that any modern autonomous vehicles misclassify, it can produce accidents. Another example is the illegal or unsuitable content can be altered in such a way that it is untraceable by the content governance algorithms used by any tool or software. Therefore, to improve the correctness of deep neural networks against adversarial attacks it is a very important step in the further advancement of the neural networks.

Chapter 2

Literature Survey

In this section, we review the techniques for attacking deep neural networks for semantic segmentation. Literature survey provided a basic foundation and analysis of knowledge on the topic. It illustrates how the proposed research is relevant to earlier research and increases statistical knowledge in the research area. Unlike image classification, there are very few researches are available on attacking semantic segmentation networks. Hence it increases the need for more research on this topic.

2.1 I-FGSM

An adversarial example image can be computed using the I-FGSM (iterative fast gradient sign method) method by adding a pixel-wide perturbation (noise) of small magnitude in the direction the same as the gradient. This perturbation is calculated iteratively, thus is efficient in terms of accuracy:

$$x_0^{adv} = x,$$
$$x_{t+1}^{adv} = x_t^{adv} + \alpha \cdot sign(\nabla_x J(x_t^{adv}, y_{true}))$$

where x is the given input (clean) image, x_{adv} is the perturbed adversarial image, J is the classification loss function, y_{true} is the actual label for the input image x. These iterative techniques take the gradient of magnitude $\alpha = \varepsilon / T$ instead of a direct single-step. One-shot methods like FGSM have lower accuracy in comparison to the iterative methods like I-FGSM in white-box attacks [10].

2.2 DAG Attack

DAG attack proposed a custom loss function, by calculating a valid set of correctly predicted target labels, called active target sets for each iteration. A large variety of adversarial examples can be generated using dense adversary generation (DAG). It also is true to various other deep

networks for segmentation and detection. It generates an adversarial perturbation (noise) given an input image and the recognition targets, which can easily confuse as many targets as possible [11].

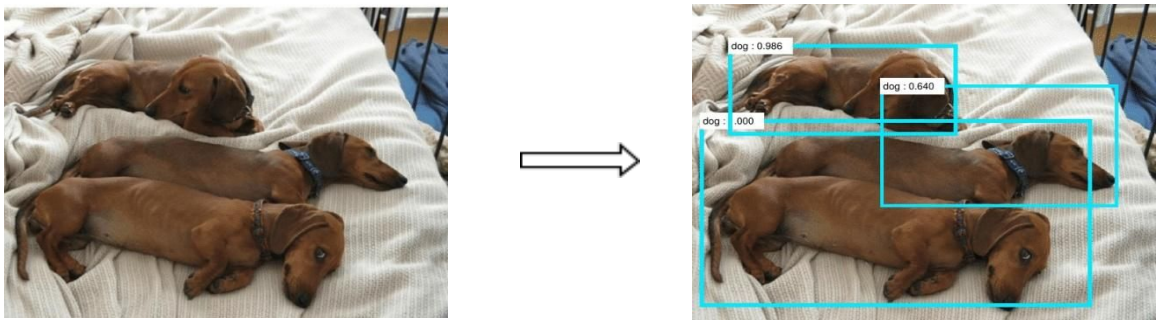


Fig 7.1 Example of object detection using Faster-RCNN on original image [11].

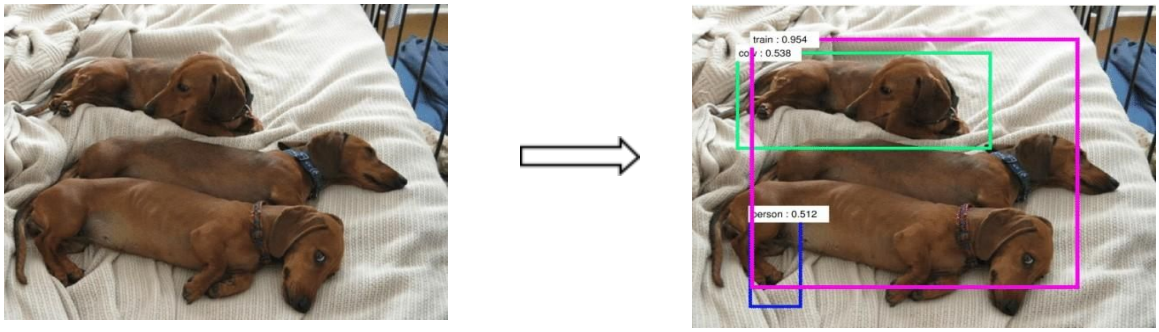


Fig 7.2 Example of object detection by Faster-RCNN on perturbation adversarial example [11].

2.3 BIM Attack

This basic iterative method presents a simple idea to generate adversarial noise. The goal is to search for a small δ so that $F(X + \delta) = y'$. The method aims to solve the following objective function:

$$\arg\min_{\delta} L(F(X+\delta), y') + c \cdot \|\delta\|_p$$

where c is responsible for the regularization of the distortion, and $||\delta||_p$ is the L_p norm that specifies $||X_{adv} - X||_p < \delta$. The optimization aims to cause a misclassification from y to y' while minimizing the perturbation to x.

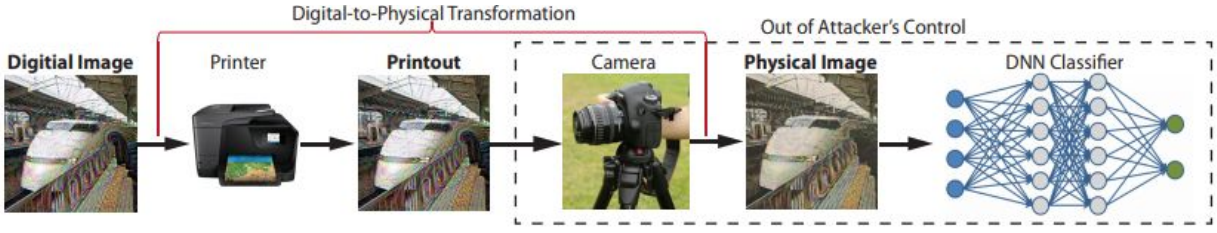


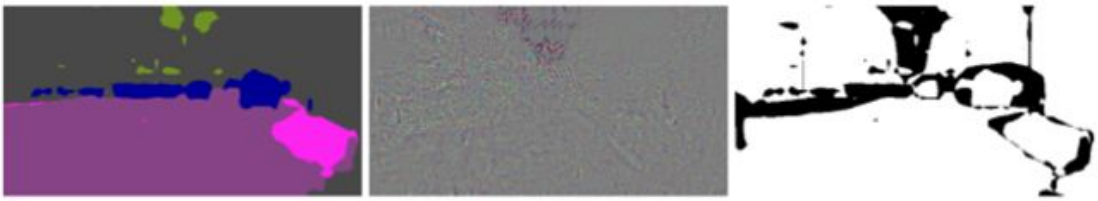
Fig 8. BIM Attack example on the digital and physical world [12].

Since the physical world challenges are not been considered by BIM. As per the above image, it is very unlikely that an attacker will directly feed a generated adversarial example (a digital image) to the classifier. Anyhow, there is a chance that a digital image can be printed by the adversary as a virtual physical object, which is then captured by the camera of the target system (for example an autonomous driving car) and digitized into a new image (referred to as "physical image"). This physical adversarial example image is the actual input of the classifier. Since the adversary has very limited control over the internal parts of the system, different angles of the picture or even non-linear response functions of the camera can affect the attack success rate.

2.4 MLAttack

MLAttack is basically a gradient-based iterative attack where we calculate the gradients iteratively to minimize the loss functions. Then we update the given input image according to these calculated gradients. The input image is changed in such a way that the labels of the generated adversarial example image display the labels of the target image. It is based on the assumption that when the actual gradients are masked by a semantic segmentation network, we can still get some significant gradient data by comparing the response of the intermediate layer source and target images. It involves the following three-stages: layer selection, gradient calculation, and generating adversarial example. In the first stage, only those intermediate layers are selected, which gives useful gradient information. Then the loss function is defined for each selected intermediate layer in the next stage. The gradient is calculated using each loss and ultimately consolidated. At last, using the consolidated gradient, this adversarial attack is performed in the final stage [13]. Some of the results of this experiment are shown in Fig 9.

(a) Source image (b) Semantic segmentation of a) (c) Targeted image (d) Semantic segmentation of c)



(e) Segmentation (f) Perturbations (g) Error image

Fig 9. Segmentation result before and after the MLAttack [13].

Chapter 3

Our Approach

3.1 Objective and Motivation

Using approximate IoU loss function to generate adversarial examples with more accuracy on any state of the art deep learning architectures for semantic segmentation. Semantic segmentation task is a basic building block of machine learning consisting of three basic steps of object detection, shape recognition, and classification. It is a very important component of autonomous driving, web security, image detection, and other computer vision tasks. So our main focus is to understand techniques to generate adversarial examples on state-of-the-art semantic segmentation models and to contribute to the same if possible.

3.2 Observations

The following observations were made during this literature survey:

- The loss function used to attack neural networks is the **cross-entropy** loss.

$$L_{CE} = -\frac{1}{N}\sum_{i=1}^{c} y_i \, log(y'_i)$$

  where y_i is the ground truth, y_i' is the predicted score, c is the total classes, and N is the number of pixels.

- Success is generally measured using Intersection-Over- Union (**IoU**), which is non-differentiable.

$$IoU = \frac{TP}{TP + FP + FN}$$

  where TP, FP, and FN refer to the counts of true positive, false positive and false negative respectively.

This **gap between performance measure and loss function** might result in a fall in performance, which has also been studied by a few recent efforts. So a method can be designed which will automatically learn a surrogate loss function, which can be used in place of cross-entropy loss. Since we have to maximize IoU, we can define a $loss_{Iou} = IoU$. Dataset processed to the format: $D = \{ Confusion\text{-}Matrix, loss_{Iou} \}$, for approximating loss function.

3.3 Neural Network Architecture

To get an approximate IoU loss function, we used a simple multilayer perceptron network. The architecture information is shown in Fig 10. It has a total of 159,775 trainable parameters. This trained model can be used to attack any segmentation model. For experiments, we used DeepLab v3 to generate adversarial examples with different configurations. DeepLab v3 [14] is among the best state-of-art deep learning models for semantic image segmentation, where it assigns semantic labels (e.g., person, car, dog, cat and so on) to every pixel in the input image. Images for this experiment are taken from PASCAL VOC 2012 [15] dataset having 1464 images for training and 1449 images for validation of dimension 513*513*3. The complete neural architecture after combining with our trained approximate IoU loss function is shown in Fig 11. After training of approximate IoU model for 20 iterations, the mean absolute error is 0.06 and the mean square error is 0.008. The results of this regression task after training are shown in Fig 12.

Layer (type)	Output Shape	Param #
flatten_1 (Flatten)	(None, 441)	0
dense_1 (Dense)	(None, 252)	111384
dropout_1 (Dropout)	(None, 252)	0
dense_2 (Dense)	(None, 126)	31878
dropout_2 (Dropout)	(None, 126)	0
dense_3 (Dense)	(None, 64)	8128
dropout_3 (Dropout)	(None, 64)	0
dense_4 (Dense)	(None, 64)	4160
dropout_4 (Dropout)	(None, 64)	0
dense_5 (Dense)	(None, 64)	4160
dropout_5 (Dropout)	(None, 64)	0
dense_6 (Dense)	(None, 1)	65

Total params: 159,775
Trainable params: 159,775
Non-trainable params: 0

Fig 10. The neural network architecture of our IoU approximation model.

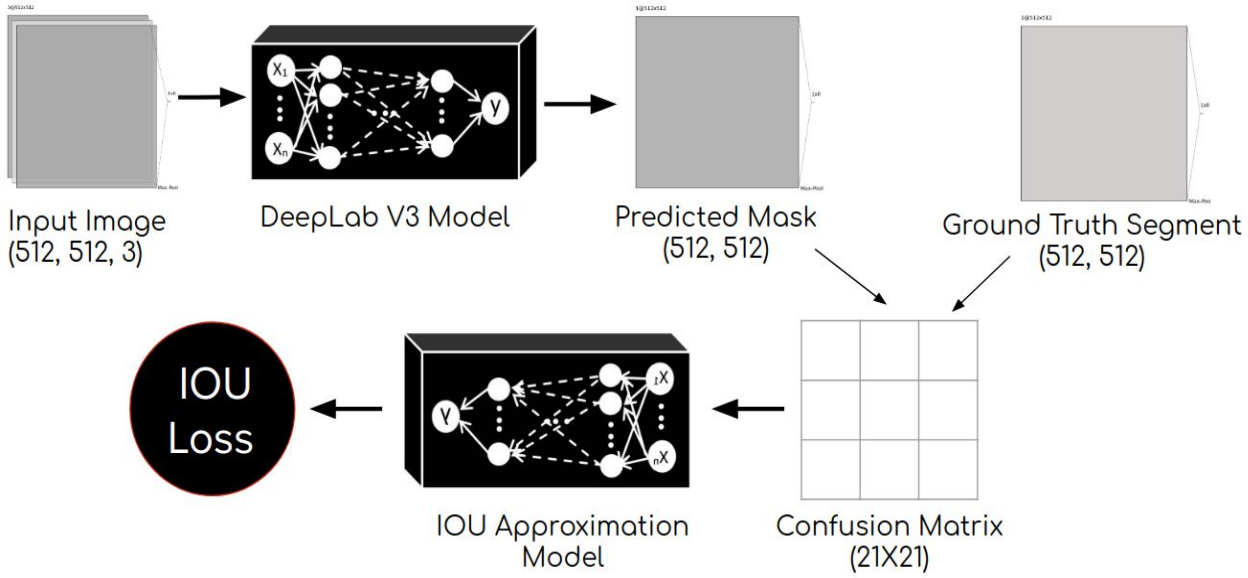
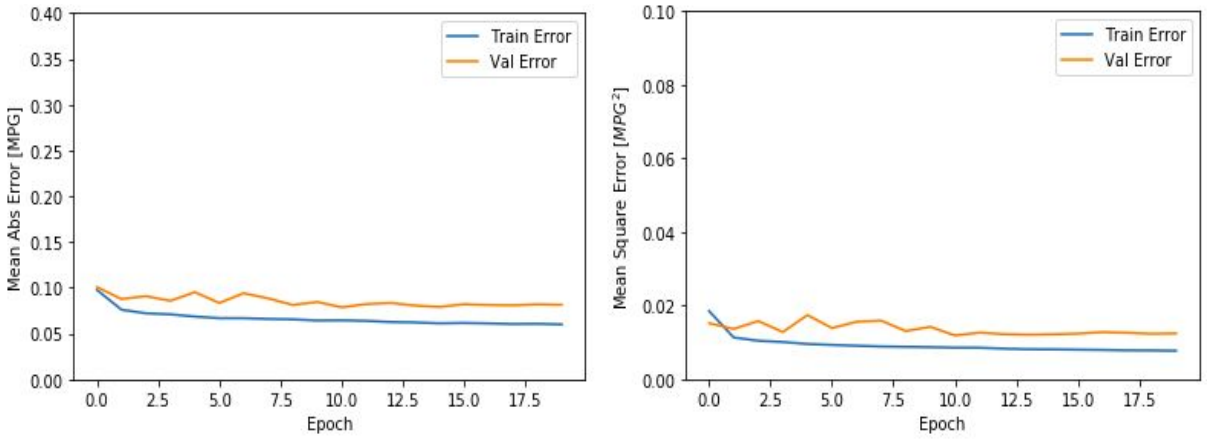


Fig 11. Complete neural network architecture after combining our trained IoU approximate model to generate adversarial examples in this experiment.

3.4 Results

3.4.1 IoU approximation model result

Training IoU approximation model is a regression task, so the mean square error was used as the loss function. The success of this method is measured using the mean absolute error and mean square error. After training for 20 iterations, the mean absolute error is 0.06 and the mean square error is 0.008 as shown in figure 12(a) and 12(b) respectively. We can clearly see that the error is decreasing with more iterations



(a) (b)

Fig 12. IoU approximation model results: (a) Plot of mean absolute error with the number of iterations (b) Plot of mean square error with the number of iterations.

3.4.2 Adversarial examples for semantic segmentation

After using approximate IoU function as loss function and changing input image while keeping weights constant during backpropagation result as shown in Fig 13 for epsilon=5/255 and in Fig 14 for epsilon=10/255. For comparison standard result from I-FGSM attack is also shown. IoU of prediction and initial segmentation is calculated and shown.

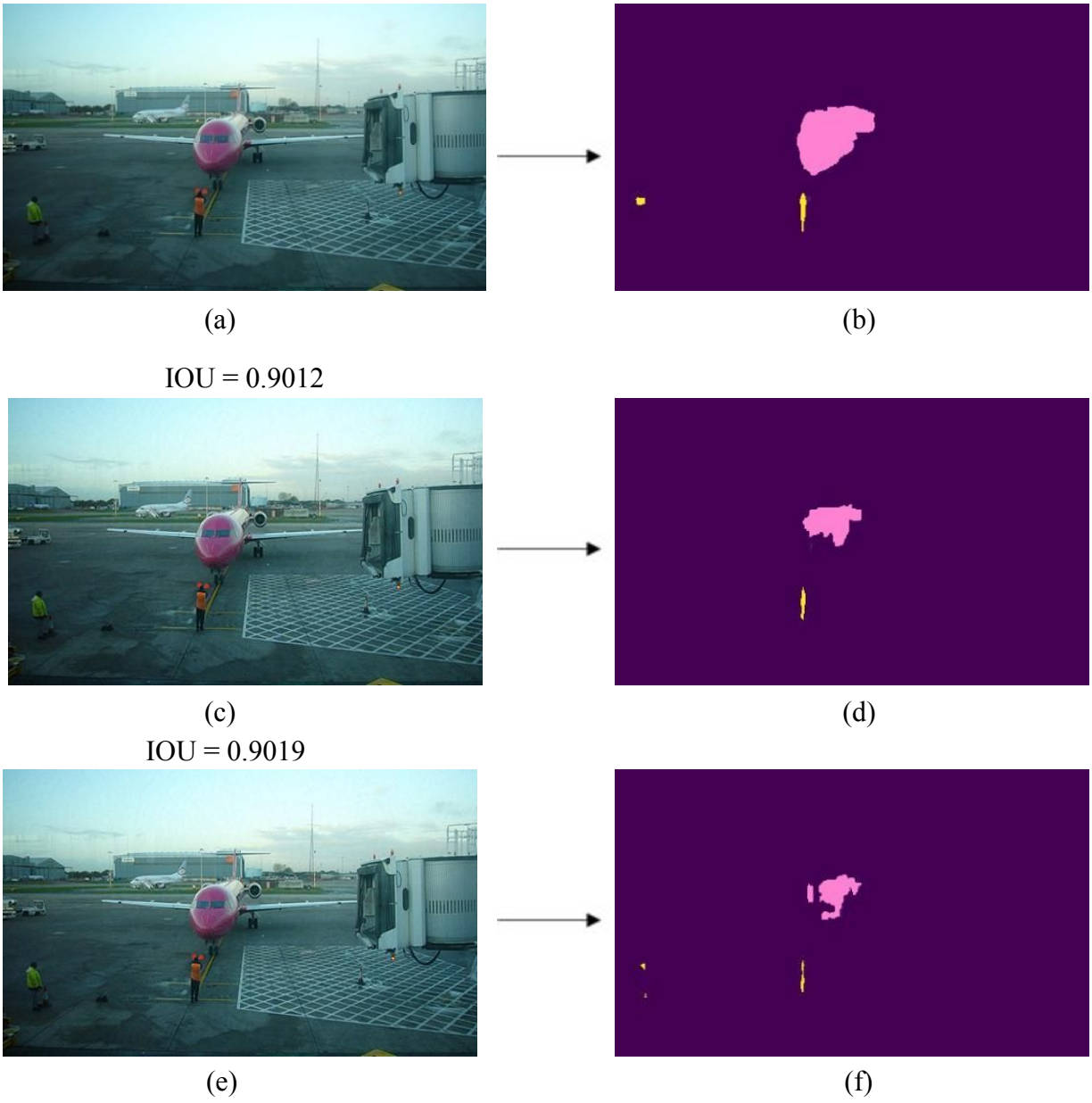
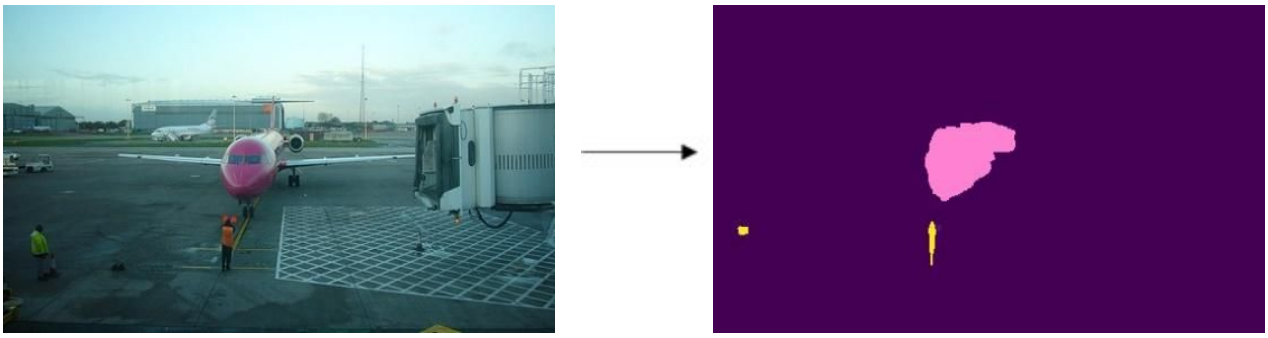
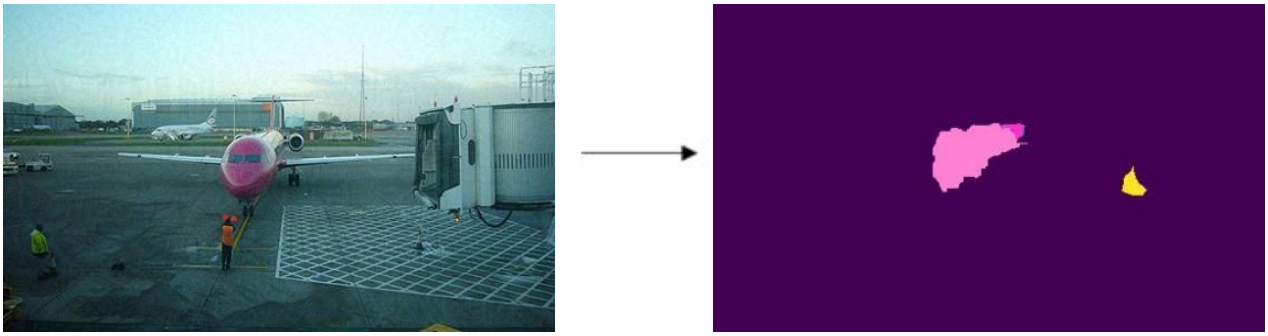


Fig 13. Results of Adversarial examples for semantic segmentation for $\epsilon = 5/255$: (a) original image (b) segmentation of original image (c) adversarial image generated using cross-entropy loss (d) segmentation of the corresponding image (e) adversarial image generated using approx. IoU loss (f) segmentation of the corresponding image.



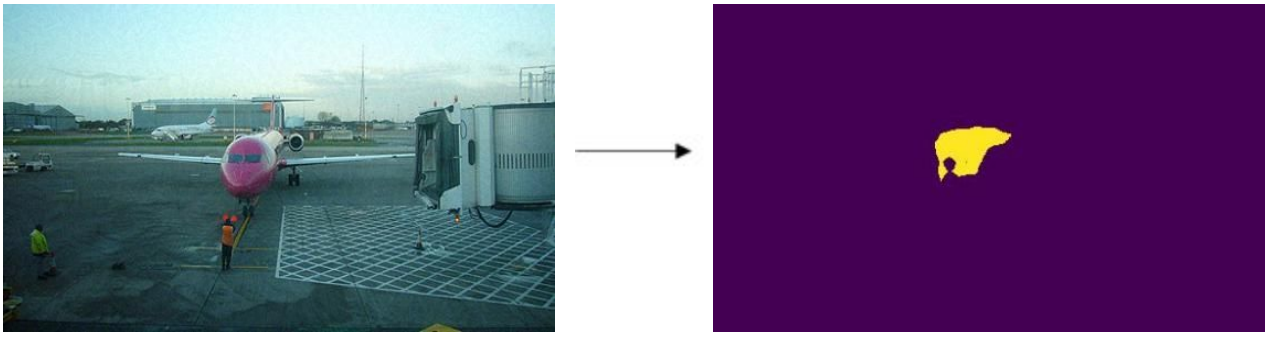
(a) (b)

IOU = 0.89169



(c) (d)

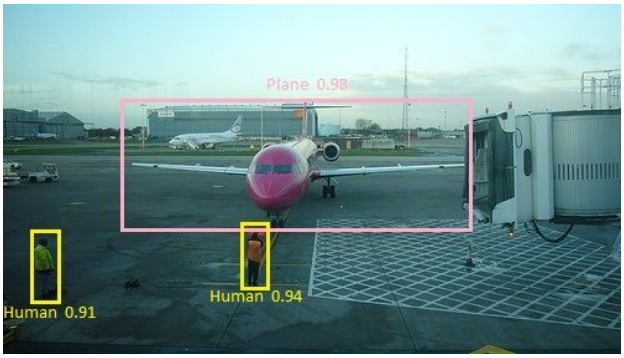
IOU = 0.911515



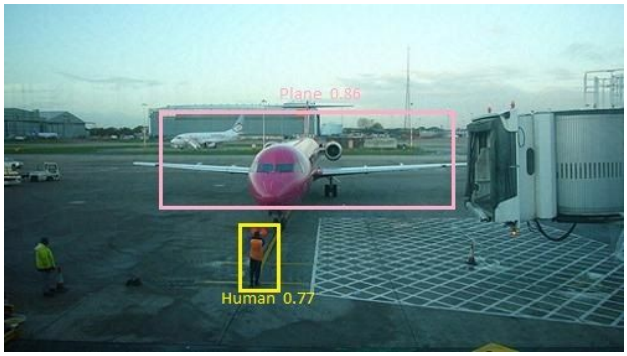
(e) (f)

Fig 14. Results of Adversarial examples for semantic segmentation for ϵ = 10/255. (a) original image (b) segmentation of the original image (c) adversarial image generated using cross-entropy loss (d) segmentation of the corresponding image (e) adversarial image generated using approx. IoU loss (f) segmentation of the corresponding image.

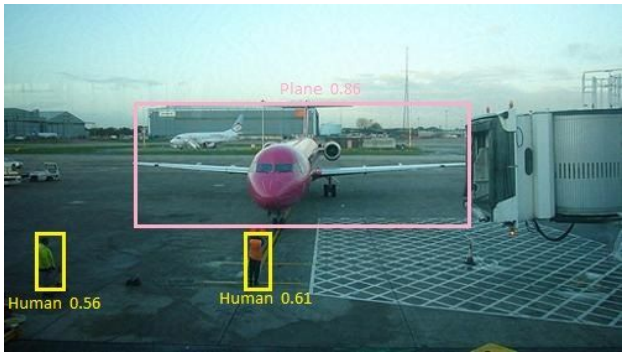
3.4.3 Adversarial examples for object detection



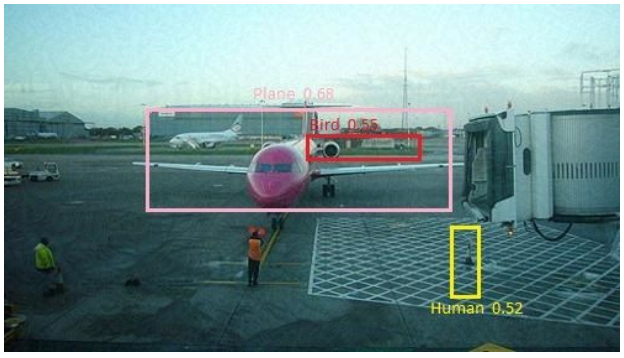
(a)



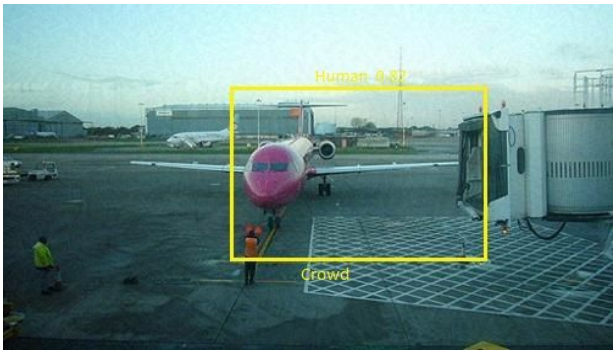
(b)



(c)



(d)



(e)

Fig 15. Object detection result: (a) original image (b) adversarial image generated using cross-entropy for ε = 5/255 (c) adversarial image generated using IoU for ε = 5/255 (d) adversarial image generated using cross-entropy for ε = 10/255 (e) adversarial image generated using IoU for ε = 10/255

Chapter 4

Conclusions and Future Work

In this report, we have tried to provide a systematic, categorical and comprehensive overview of the recent works and researches related to adversarial attacks. We have established one really good, and intuitive method of using approximate IoU loss function to generate adversarial examples. We started with basic IoU approximation using a simple neural network and solved the problem of non-differentiability. The proposed attack has successfully produced the adversarial examples to trick the well known semantic segmentation network (Deeplab v3) and made them predict the incorrect segment. We have cross verified our attack image on one object detection technique and it worked excellent there too. This trained approximate model can also be used with any attack technique to produce adversarial examples. The current findings suggest that a custom approximate IoU loss function could outperform basic adversarial attacks for semantic segmentation, given that approximate neural network is trained very well.

In the future, we will improve the results by improving the approximate IoU loss function by increasing training data or changing the hyperparameter of the neural network. Alternatively, implementing an ensemble method by combining both the loss functions will further improve results. We will work on the optimization of our work and also try to develop support for different deep learning semantic segmentation neural network models. Moreover, the current attack technique can also be used for targeted attacks.

References

1. Qi Wang, Meihan Wu, Fei Yu, Chen Feng *, Kaige Li, Yuemei Zhu, Eric Rigall and Bo He. “Real-Time Semantic Segmentation Network for Side-Scan Sonar Images”. Sensors 19, no. 9, 2019.
   https://www.mdpi.com/1424-8220/19/9/1985/review_report
2. “Tricking a Machine into Thinking You’re Milla Jovovich”(On Medium, 9 Aug 2019).
   https://medium.com/element-ai-research-lab/tricking-a-machine-into-thinking-youre-milla-jovovich-b19bf322d55c
3. Xiaoyong Yuan, Pan He, Qile Zhu, Xiaolin Li. “Adversarial Examples: Attack and Defenses for Deep Learning”. IEEE transactions on neural networks and learning systems, Jan 2019.
   https://arxiv.org/abs/1712.07107
4. Mahmood Sharif, Sruti Bhagavatula. “Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition”. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, ACM, 2016.
   https://dl.acm.org/citation.cfm?id=2978392
5. Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, Dawn Song. “Robust Physical-World Attacks on Deep Learning Models”. Conference on Computer Vision and Pattern Recognition, 2018
   https://arxiv.org/abs/1707.08945
6. “Intersection over Union (IoU) for object detection”(November 7, 2016). Adrian Rosebrock.
   https://www.pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/
7. Taga, K, Kameyama, K.; Toraichi, K. “Regularization of hidden layer unit response for neural networks”. IEEE Pacific Rim Conference on Communications Computers and Signal Processing, Vol. 1, Aug 2003
   https://ieeexplore.ieee.org/document/1235788

8. Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, Ananthram Swami. “Practical Black-Box Attacks against Machine Learning”. ACM on Asia conference on computer and communications security, Apr 2017. https://dl.acm.org/citation.cfm?id=3053009
9. Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy. “Explaining and Harnessing Adversarial Examples”. arXiv preprint arXiv:1412.6572, Dec, 2014. https://arxiv.org/abs/1412.6572
10. “Adversarial Attacks and Defences for Convolutional Neural Networks”. (Medium) https://medium.com/onfido-tech/adversarial-attacks-and-defences-for-convolutional-neural-networks-66915ece52e7
11. Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, Alan Yuille. “Adversarial Examples for Semantic Segmentation and Object Detection”. IEEE International Conference on Computer Vision 2017. https://arxiv.org/abs/1703.08603
12. Steve T.K. Jan1, Joseph Messou, Yen-Chen Lin, Jia-Bin Huang, and Gang Wang. “Connecting the Digital and PhysicalWorld: Improving the Robustness of Adversarial Attacks”. Article, Researchgate, July 2019. https://www.researchgate.net/publication/335800839_Connecting_the_Digital_and_Physical_World_Improving_the_Robustness_of_Adversarial_Attacks
13. Puneet Gupta, Esa Rahtu. “MLAttack: Fooling Semantic Segmentation Method by Multi-Layer Attack.” In German Conference on Pattern Recognition, Springer, Cham, Sep 2019. https://www.springerprofessional.de/en/mlattack-fooling-semantic-segmentation-networks-by-multi-layer-a/17315850
14. DeepLab: Deep Labelling for Semantic Image Segmentation. https://github.com/tensorflow/models/tree/master/research/deeplab
15. Visual Object Classes Challenge 2012 (VOC2012). http://host.robots.ox.ac.uk/pascal/VOC/voc2012

16. G Nagendar, Digvijay Singh, V. Balasubramanian, C.V Jawahar. “Neuro-IoU: Learning a Surrogate Loss for Semantic Segmentation”. In BMVC, Sep 2018. http://bmvc2018.org/contents/papers/1055.pdf
17. Anurag Arnab, Ondrej Miksik, Philip H.S. Torr. “On the Robustness of Semantic Segmentation Models to Adversarial Attacks”. IEEE Conference on Computer Vision and Pattern Recognition, 2018. https://arxiv.org/abs/1711.09856